DOCUMENT RESUME

ED 387 515 TM 023 748

AUTHOR Williams, Valerie S. L.
TITLE Some Advantages of Controlling for False Discoveries in Multiple Comparisons.
PUB DATE 19 Apr 95
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Hypothesis Testing; Simulation; *Statistical Analysis
IDENTIFIERS Bonferroni Procedure; Error Detection; *False Discoveries; *Multiple Comparisons; Power (Statistics); Type I Errors

ABSTRACT

Multiple comparison procedures for controlling familywise Type I error and the false discovery rate are described and compared, including the traditional Bonferroni correction, a sequential (step-up) Bonferroni procedure (Hochberg, 1988), and a sequential false discovery rate procedure proposed by Benjamini and Hochberg (1995). Motivation for formally considering the false discovery rate is discussed. A simulation study demonstrates that the Benjamini and Hochberg (BH) technique results in greater power than either of the Bonferroni procedures, and the power advantage increases with the number of inferences in the family. Another important advantage of the BH procedure is its relative consistency about the statistical significance of comparisons over alternative family sizes. It is concluded that, in situations where there is a great number of hypotheses to be tested and strong control of familywise error is unnecessary, it is reasonable to apply the BH technique as a statistical approach to error control. (Contains 12 references, 2 tables, 2 figures, and an appendix of Statistical Analysis System code for implementing the Bonferroni and BH procedures.) (Author)

Some Advantages of Controlling for False Discoveries in Multiple Comparisons

Valerie S. L. Williams
National Institute of Statistical Sciences

ABSTRACT

Multiple comparison procedures for controlling familywise Type I error and the false discovery rate are described and compared, including the traditional Bonferroni correction, a sequential (step-up) Bonferroni procedure (Hochberg. 1988), and a sequential false discovery rate procedure proposed by Benjamini and Hochberg (1995). Motivation for formally considering the false discovery rate is discussed. A simulation study demonstrates that the Benjamini and Hochberg (BH) technique results in greater power than either of the Bonferroni procedures, and the power advantage increases with the number of inferences in the family. Another important advantage of the BH procedure is its relative consistency about the statistical significance of comparisons over alternative family sizes. It is concluded that, in situations where there is a great number of hypotheses to be tested and strong control of familywise error is unnecessary, it is reasonable to apply the BH technique as a statistical approach to error control.

Please address correspondence to: Valerie S. L. Williams, *National Institute of Statistical Sciences*, P.O. Box 14162, Research Triangle Park, NC 27709-4162.
Email: *williams@niss.rti.org*

Paper presented at the Annual Meeting of the
American Educational Research Association,
San Francisco, CA (April 19, 1995)

Some Advantages of Controlling for False Discoveries in Multiple Comparisons

For any study, it is necessary to use statistical procedures to maintain some kind of error rate when testing a statistical hypothesis. By convention, the Type I error rate for the test of a single hypothesis, such as H_0: $\mu_1 - \mu_2 = 0$, is usually set to $\alpha = .05$; that is, there is a 5% probability that a hypothesis will be rejected when, in fact, it is true.

It is very uncommon for an educational researcher to conduct a survey or design an experiment to measure only one variable and test only one hypothesis in order to answer a single question. Multiplicity, and the need for multiple comparison procedures, arises in the more common situation where more than one statistical hypothesis is evaluated. Unless some correction is incorporated, the simultaneous or familywise Type I error rate — the probability of one or more false rejections in a given set or family of hypotheses — will exceed the nominal α (which actually applies to any single test considered alone).

Definitions

Let p_{crit} be the critical p-value associated with the null sampling distribution of the test statistic for any hypothesis test. Let m be the number of comparisons and $i = 1, \ldots, m$ be the rank of the observed p-value, $p_{(i)}$, ordered from smallest to largest, so that they are weakly increasing from $i = 1$ to $i = m$.

(1) The unadjusted test or per comparison approach: Declare any comparison statistically significant if

$$p_{(i)} \leq p_{crit} = p_{UN} = \alpha/2 .$$

(2) The traditional Bonferroni correction is a simple and trustworthy statistical procedure for assuring simultaneously that the probability of at least one Type I error is no greater than α. Declare any comparison statistically significant if

$$p_{(i)} \leq p_{crit} = p_{B1} = \alpha/2m .$$

Two sequential approaches are defined as follows:

(3) The step-up Bonferroni procedure (Hochberg, 1988): Declare the i^{th} comparison statistically significant when, beginning with $i = m$ and continuing toward $i = 1$,

$$p_{(i)} \leq p_{crit} = p_{B2}(i) = \alpha/2(m-i+1)$$

then stop and declare significance for all comparisons for which $j \leq i$.

(4) The Benjamini and Hochberg (1995) procedure (BH) controls the false discovery rate or the average proportion of declared significances which are erroneous: Declare the i^{th} comparison statistically significant when, beginning with $i = m$ and continuing toward $i = 1$,

$$p_{(i)} \leq p_{crit} = p_{BH}(i) = i\alpha/2m$$

then stop and declare significance for all comparisons for which $j \leq i$.

Both the step-up Bonferroni and the BH procedure are sequential multiple comparison techniques which provide greater statistical power than the simple Bonferroni correction while still attempting to control the overall error rates. The BH procedure controls familywise error in the weak sense, that is, Type I error is bounded by α only in the complete null case, when all null hypotheses are true. The step-up Bonferroni adjustment controls familywise error in the strong sense, that is, Type I error is bounded by α under all configurations of hypotheses; however, it is known to be conservative and lacking in statistical power, especially in the case of very large family sizes. Both techniques provide "strong" control of the false discovery rate.

An Illustration from the NAEP TSA

Summary data from the National Assessment of Educational Progress (*NAEP*) Trial State Assessment (*TSA*) are used to illustrate the multiple comparison procedures. The data are average eighth-grade mathematics scores for the 34 states which participated in both the 1990 and 1992 *NAEP TSA* (Johnson, Mazzeo, & Kline, 1993). Table 1 contains the computed differences between 1990 and 1992 mathematics achievement means and the computed pooled standard errors. The table also includes the decision about statistical significance, indicated by an "*" in each column, for each state's change under the four multiplicity treatments ($\alpha = .05$).

The unadjusted per comparison approach (p_{UN}) finds 15 differences to be statistically significant, whereas the BH procedure (p_{BH}) finds 11. An ordinary Bonferroni adjustment for control of Type I error, as used by *NAEP*, results in a critical *p*-value of $p_{B1} = .000735$ — the simple Bonferroni correction (p_{B1}) indicates only four significant differences, as does the step-up Bonferroni (p_{B2}) correction.

Insert Table 1 about here.

Mean (and standard error) eighth-grade mathematics achievement change by state, 1990 to 1992, *t*, *p*-value, and p_{crit}-values for four multiple comparison adjustments. $m = 34$ ($df \approx 60^{\dagger}$).

As a further example, all pairwise mean differences between the states' 1992 eighth-grade mathematics achievement scores were also compared. There were 41 states which participated in the 1992 assessment, resulting in a family size of $m = 41 \times 40/2 = 820$. Table 2 summarizes the number of statistically significant differences among the means. By the Bonferroni adjustment (using a critical *p*-value of $p_{B1} = .0000304$), there are 480 significant differences between pairs of states; the step-up Bonferroni admits 13 more rejections, and the use of the BH procedure results in an additional 159 declarations of significance. The unadjusted analysis increases the number of statistically significant differences beyond the BH technique by only 6.

Insert Table 2 about here.

Number of statistically significant differences between all pairs of states, $m = 820$ ($df \approx 60$).

Three Advantages of Controlling for False Discoveries

There are three primary advantages of applying a technique, such as the BH procedure, for the control of the rate of false discoveries in multiple hypothesis testing:

- formal consideration of an alternative error rate,
- increased statistical power, and
- consistency of findings over differing family sizes.

Each of these will be considered in turn.

Formal Consideration of an Alternative Error Rate

The traditional emphasis on control of the Type I error rate and familywise error neglects the practical importance of failure to detect true differences, i.e., committing a Type II error. Although familywise Type I error rate is very easily controlled by simply manipulating the value of α, Hays (1988) recommends that "in some situations, perhaps, we should be far more attentive to Type II errors, and less attentive to setting α at one of the conventional levels" (p. 263). Many others call for placing a higher priority on ability to detect real differences (cf. Carmer & Swanson, 1973; Soric, 1989).

As Tamhane (1995) points out, familywise control of Type I error is essential in studies "where the correctness of an overall decision depends on the simultaneous correctness of all individual inferences" (p. 3); however, there are situations where "one erroneous comparison will generally not jeopardize the conclusion" (Benjamini, Hochberg, & Kling, 1995, p. 6). Moreover, in some circumstances where a set of statistics are evaluated, an overall decision may even be unnecessary and, instead, many separate decisions and recommendations are involved — in such cases, it may be more worthwhile to be assured of the rate of false discovery rather than the probability that all inferences are correct.

Shaffer (1994), too, notes that false discovery rate control may be an attractive alternative to strict control of the familywise Type I error rate; for example, when examining all possible differences among pairs, researchers may be willing to admit a certain small proportion of errors in order to discover as many significant differences as possible.

Statistical Power: A Simulation Study

To investigate the statistical power of the adjustment techniques, a simulation study was performed, and structured to be similar to the data from the *NAEP TSA*. For each of 48 states, mean "achievement levels," μ_i, were defined to be the approximate median values of each of 48 ordered random observations from a normal $(0,\sigma_A)$ distribution (for which $s^2 = .98$). Five conditions of effect size were studied by setting the value of σ_A. For the near-null condition of negligible differences among the μ_i, the value of σ_A was set to 0.001; four non-

null conditions were considered, with increasing values of σ_A: {0.3, 1.0, 3.0, and 5.0}. In each case, for each of 10,000 replicates, an observed mean for each state, $\bar{X}_i$, was generated by adding a number randomly selected from a normal (0,1) distribution to the corresponding μ_i.

In the first of two families studied, each $\bar{X}_i$ was compared to a "national mean," treated here as a known constant, M. This results in $m = 48$ independent comparisons about which we wish to determine the significance of $\mu_i - M$. The second family was comprised of all pairwise comparisons where each $\bar{X}_i$ was compared with each $\bar{X}_j$, resulting in $m = 1128$ comparisons about which we wish to determine the significance of $\mu_i - \mu_j$. (See Williams, Jones, and Tukey (1994) for a more detailed description.)

Figure 1 presents plots of the statistical power against effect size for the BH and the two Bonferroni adjustment techniques, B1 and B2, $\alpha = .05$. Power is defined as what Hochberg and Tamhane (1987) refer to as *all-pairs power*, the probability of claiming significance for all true differences among all pairs; it is calculated as the average proportion of rejections claimed over the 10,000 replications.

Insert Figure 1 about here.

Average statistical power for the BH and Bonferroni (B1 and B2) techniques,
48 independent comparisons (above) and 1128 pairwise differences among the 48 (below).

Under all effect-size conditions, for the independent and pairwise comparisons families, the BH technique results in greater power than that for either the conventional Bonferroni or the step-up Bonferroni procedures. The relative advantage in power for the BH technique is greatest for the large pairwise family and for large effect sizes. These results are consistent with the simulation findings reported by Benjamini and Hochberg (1995) and Benjamini, Hochberg, and Kling (1994).

Two more sets of data were simulated to examine the effects of partial dependence and family size on statistical power. In one condition, each of 1128 values of $\bar{X}_i$ was compared to M, the "national mean," yielding a family of independent comparisons of the same size as the family of pairwise differences above. In the second condition, 10 state mean values were compared among themselves as differences among all possible pairs ($m = 10 \times 9/2 = 45$), resulting in a family size similar to that for the 48 independent comparisons above. The same five conditions of effect size were studied, with $\alpha = .05$.

Figure 2 presents further plots of statistical power against effect size for the adjustment techniques. The BH technique results in greater power than the other procedures under all effect-size conditions, for both the 1128 independent comparisons (upper plot) and the pairwise comparisons among 10 (lower plot). The relative advantage in power for the BH technique is greatest for the large effect sizes. Comparing the results in Figure 2 with those presented in

Figure 1, it is clear that the BH advantage in power is associated with the large family size and is little affected by the dependence or the independence of the contrasts tested.

Insert Figure 2 about here.

Average statistical power for the BH and Bonferroni (B1 and B2) techniques, 1128 independent comparisons (above) and 45 pairwise differences among 10 (below).

Consistency of Findings over Differing Family Sizes

Saville (1990) makes a strong case for eschewing altogether multiple comparison procedures on the grounds that they frequently lead to inconsistent decisions regarding statistical significance. Multiple comparison procedures are inconsistent when the significance of a given contrast value with a given standard error can vary from "not significant" to "highly significant," depending simply on family size. Family size, or m, is defined as the number of inferences under consideration; however, there are often legitimate ambiguities about what constitutes *the* family of interest for a particular set of data. In large-scale comparative education studies, such as *NAEP*, the determination of family size is critical to the implementation of a multiple comparison procedure.

Using a real data example from *NAEP*, Williams, Jones, and Tukey (1994) demonstrated that the BH procedure was a relatively consistent multiplicity adjustment across several different plausible definitions of family. The conventional Bonferroni correction was the least consistent and the most conservative, and the step-up Bonferroni procedure performed very similarly. In this particular example, the BH procedure with the most conservative definition of family admitted more statistically significant differences than the conventional Bonferroni applied to the smaller, more lenient, family sizes.

The general lack of invariance of the various multiple comparison techniques is also suggested in Table 2, where family size is defined as $m = 820$, all possible pairwise comparisons among 41 states. The unadjusted result of 658 statistical significances will always be consistent (logically, when no adjustment is made, there are no inconsistencies because each comparison essentially consists of a unique family of size $m = 1$, and $p_{crit} = .05$ is applied throughout, regardless of the total number of simultaneous inferences). For the simple Bonferroni and the step-up Bonferroni techniques, there are 480 and 493 significant differences, respectively. The BH procedure is comparable to the unadjusted per comparison method as it allows 652 rejections. If family size were redefined to consider, say, 41 separate families of 40 comparisons (each state compared with each other state), there would again be 658 statistical significances from the unadjusted approach, and somewhat more (up to a maximum of 658) for each of the other adjustment procedures because of the increase in p_{crit} associated with smaller m.

Conclusions

Although the BH and step-up Bonferroni techniques are both sequential in nature, they are easily implemented (see the Appendix for SAS code).

A desirable feature for a multiple comparison procedure is that it provide decisions about significance that are relatively invariant over alternative choices of family size. The BH technique behaves more consistently than the more conservative Bonferroni techniques so that discrepancies in the reporting of different statistical results to different audiences with different interests are, to some extent, minimized.

The BH procedure demonstrates an important gain in power over the simple Bonferroni and step-up Bonferroni adjustments. As indicated from the results shown both here and in Benjamini and Hochberg's (1995) simulation studies, the power advantage of the BH procedure increases with the number of comparisons when the true differences remain about the same size: The loss of power with increasing m for the BH technique is slower than the corresponding loss of power for the Bonferroni adjustment. This may seem too good to be true, but it is important to keep in mind that the error rate controlled here is different: The BH procedure maintains a 5% error rate such that only 1 out of 20 declarations of significance are erroneous; the conservatism of the Bonferroni procedures is due to the small p_{crit} required for strong protection against Type I error.

The traditional emphasis on control of the Type I error rate and familywise error neglects the importance of failing to detect true differences, the Type II error. Recalling that power is the complement of the Type II error rate — power = 1 - β, where β is the Type II error rate — it is apparent that in the basic configurations presented in the simulation study the BH procedure results in far fewer Type II errors than do either of the Bonferroni techniques. The error control of the BH multiple comparison procedure can be characterized as providing strong control of the false discovery rate and weak control of the familywise Type I error rate — as the configuration of the set of hypotheses moves away from the general null case, the control of familywise error (and the false discovery rate) gives way to control of the false discovery rate. The BH procedure is an approach worth considering whenever it is acceptable to entertain the particular redefinition of α that the BH procedure invokes.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289-300.

Benjamini, Y., Hochberg, Y., & Kling, Y. (1995). False discovery rate controlling procedures for pairwise comparisons. Unpublished manuscript.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association, 68*, 66-74.

Hays, W. L. *Statistics, 4th edition*. Fort Worth, TX: Holt, Rinehart, and Winston.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803.

Johnson, E. G., Mazzeo, J., & Kline, D. L. (1993). *Technical report of the NAEP 1992 Trial State Assessment program in mathematics*. Report 23-ST05. Washington, DC: National Center for Education Statistics.

Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *The American Statistician*, 44, 174-180.

Shaffer, J. P. (1994). *Multiple hypothesis testing: A review*. (Technical Report #23). Research Triangle Park, NC: National Institute of Statistical Sciences.

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association, 84*, 608-610.

Tamhane, A. (1995). Multiple comparison procedures. Unpublished manuscript.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1994). *Controlling error in multiple comparisons, with special attention to the National Assessment of Educational Progress* (Technical Report #33). Research Triangle Park, NC: National Institute of Statistical Sciences.

Table 1.

Mean (and standard error) eighth-grade mathematics achievement change by state, 1990 to 1992, t, p-value, and p_{crit}-values for four multiple comparison adjustments, $m = 34$ (df taken as 60†).

State	$\bar{X}_{92}-\bar{X}_{90}$ (se)	t	(p-value)	p_{UN}	$p_{BH}(i)$	$p_{B2}(i)$	p_{B1}
GA	-0.323 (1.77571)	-0.18190	(.42814)	.025	.025000	.025000	.000735
AR	-0.777 (1.48529)	-0.52313	(.30141)	.025	.024265	.012500	.000735
AL	-1.568 (2.01745)	-0.77722	(.22004)	.025	.023529	.008333	.000735
NJ	1.565 (1.92728)	0.81203	(.20999)	.025	.022794	.006250	.000735
NE	1.334 (1.52772)	0.87320	(.19320)	.025	.022059	.005000	.000735
ND	1.526 (1.68552)	0.90536	(.18445)	.025	.021324	.004167	.000735
DE	1.374 (1.34651)	1.02042	(.15581)	.025	.020588	.003571	.000735
MI	2.215 (1.84727)	1.19906	(.11761)	.025	.019853	.003125	.000735
LA	2.637 (2.07943)	1.26814	(.10482)	.025	.019118	.002778	.000735
IN	2.149 (1.63556)	1.31392	(.09694)	.025	.018382	.002500	.000735
WI	2.801 (1.96269)	1.42713	(.07936)	.025	.017647	.002273	.000735
VA	2.859 (1.92992)	1.48141	(.07187)	.025	.016912	.002083	.000735
WV	2.331 (1.39639)	1.66930	(.05013)	.025	.016176	.001923	.000735
MD	3.399 (1.92320)	1.76737	(.04113)	.025	.015441	.001786	.000735
CA	3.777 (2.11460)	1.78615	(.03956)	.025	.014706	.001667	.000735
OH	3.466 (1.85022)	1.87329	(.03295)	.025	.013971	.001563	.000735
NY	4.893 (2.53195)	1.93250	(.02901)	.025	.013235	.001471	.000735
PA	4.303 (2.20545)	1.95108	(.02786)	.025	.012500	.001389	.000735
FL	3.784 (1.93266)	1.95792	(.02745)	.025	.011765	.001316	.000735
WY	2.226 (1.09641)	2.03026	(.02339)	.025 *	.011029	.001250	.000735
NM	2.334 (1.14816)	2.03282	(.02325)	.025 *	.010294	.001190	.000735
CT	3.204 (1.53443)	2.08807	(.02052)	.025 *	.009559	.001136	.000735
OK	4.181 (1.75467)	2.38278	(.01018)	.025 *	.008824	.001087	.000735
KY	4.327 (1.61804)	2.67422	(.00482)	.025 *	.008088 *	.001042	.000735
AZ	4.994 (1.85110)	2.69785	(.00452)	.025 *	.007353 *	.001000	.000735
ID	2.956 (1.06775)	2.76845	(.00374)	.025 *	.006618 *	.000962	.000735
TX	5.645 (1.88770)	2.99041	(.00202)	.025 *	.005882 *	.000926	.000735
CO	4.326 (1.38868)	3.11519	(.00141)	.025 *	.005147 *	.000893	.000735
IA	4.811 (1.48805)	3.23309	(.00100)	.025 *	.004412 *	.000862	.000735
NH	4.422 (1.35399)	3.26591	(.00090)	.025 *	.003676 *	.000833	.000735
NC	7.265 (1.58701)	4.57779	(.00001)	.025 *	.002941 *	.000806 *	.000735 *
HI	5.550 (1.17134)	4.73817	(.00001)	.025 *	.002206 *	.000781 *	.000735 *
MN	6.421 (1.35226)	4.74836	(.00001)	.025 *	.001471 *	.000758 *	.000735 *
RI	5.097 (0.94844)	5.37407	(.00000)	.025 *	.000735 *	.000735 *	.000735 *
t_{crit}				2.00	2.47	3.30	3.33

* Confident direction of change.

† *Note*: The number of students sampled within states is generally close to 2000; however, because of the clustered nature of the sample design and the use of plausible values in *NAEP*, the effective sample size per state is estimated to be about 30, so that the degrees of freedom for a pairwise mean comparison is about 60.

Table 2.

Number of statistically significant differences between all pairs of states, $m = 820$ ($df \approx 60$).

Procedure	Number
Unadjusted	658
BH	652
Step-up Bonferroni	493
Bonferroni	480

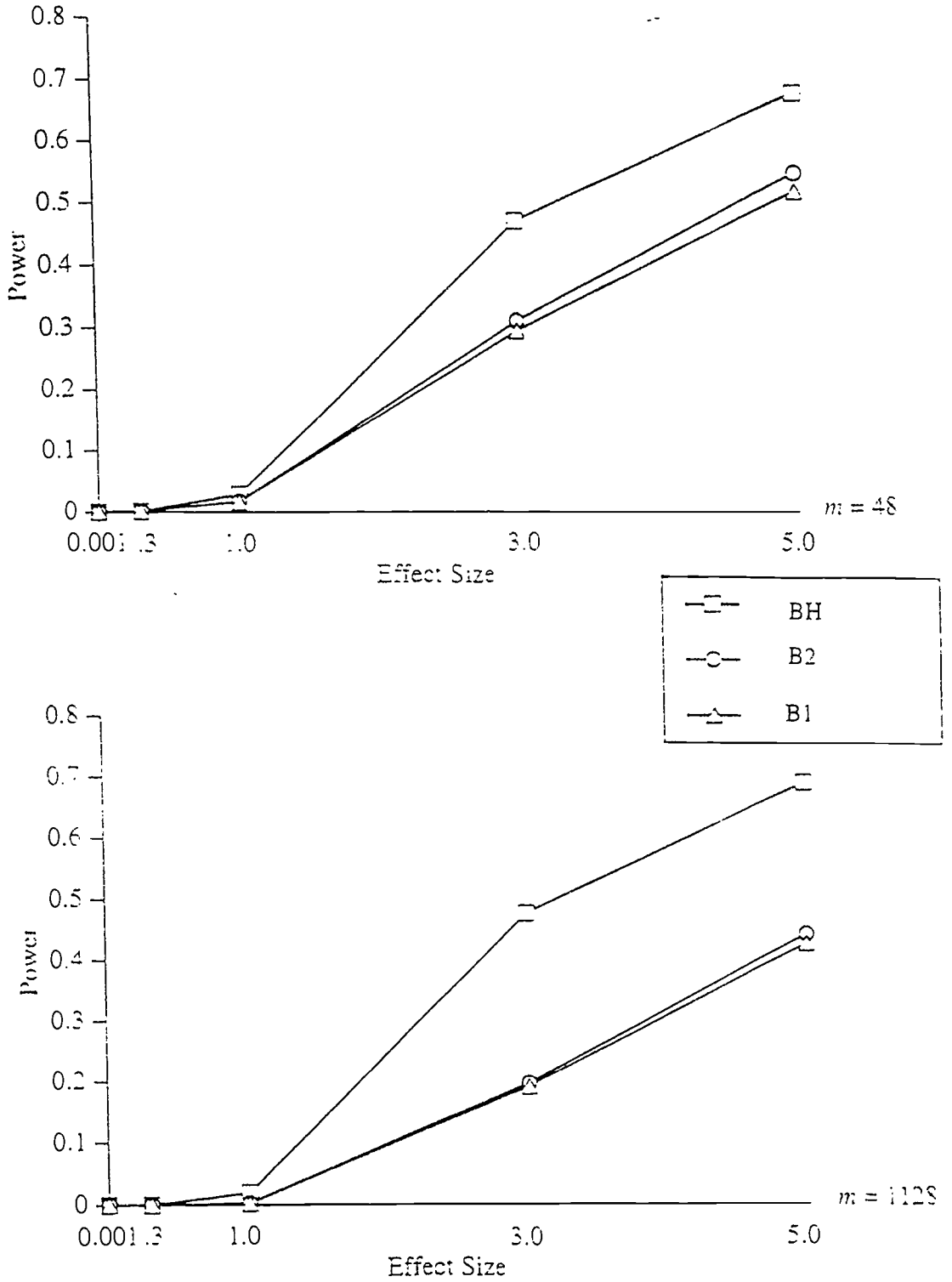

Figure 1.

Average statistical power for the Benjamini and Hochberg procedure, and the simple and step-up Bonferroni techniques, 48 independent differences (above) and 1128 pairwise differences among the 48 (below).

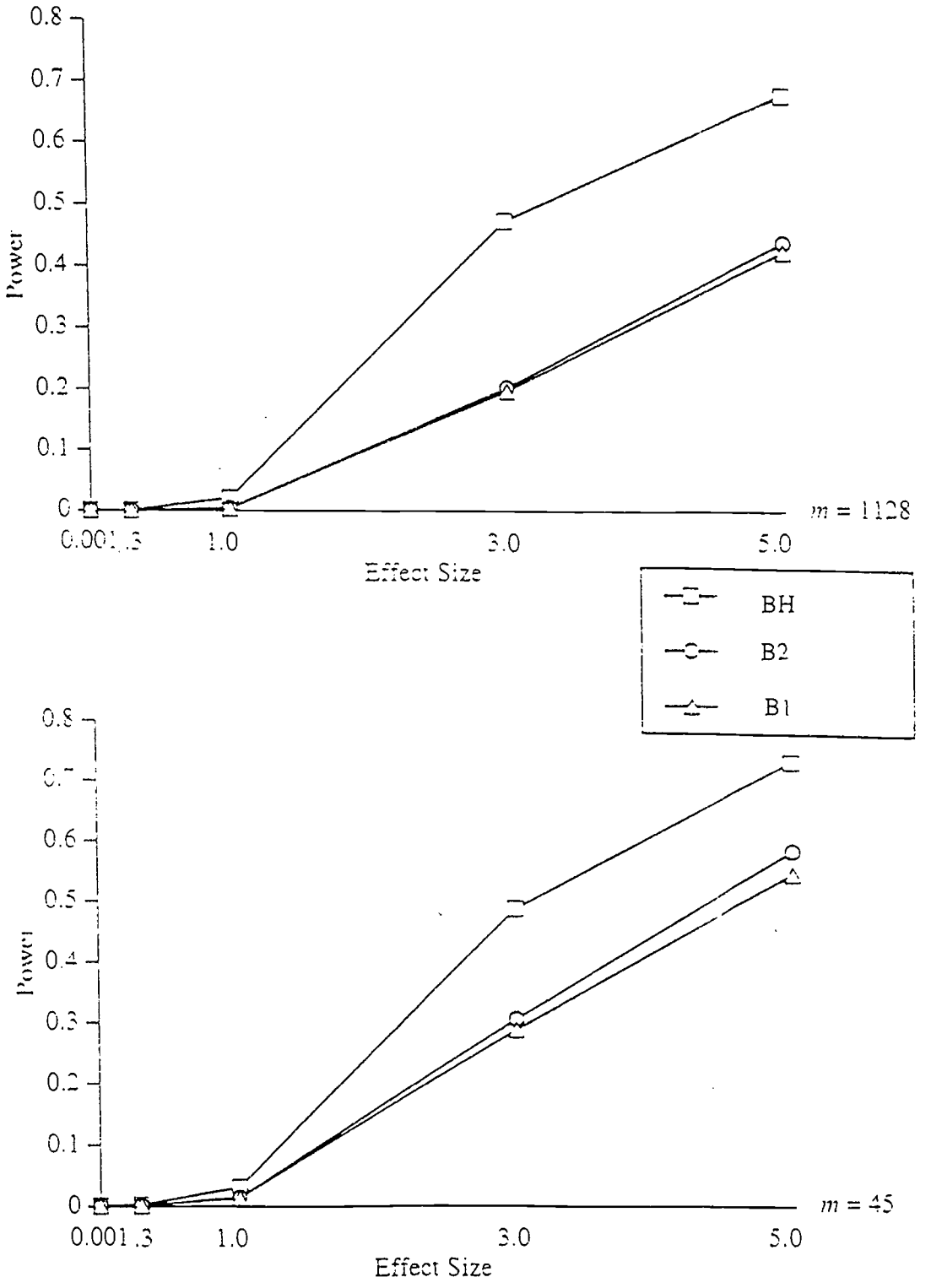

Figure 2.

Average statistical power for the Benjamini and Hochberg procedure, and the simple and step-up Bonferroni techniques, 1128 independent differences (above) and 45 pairwise differences among 10 (below).

Appendix

SAS Code for Implementing

the Step-Up Bonferroni and the BH Multiple Comparison Procedures

```
*********************************************************************
Computes t-statistics evaluating the state change compared to the
 average change for NAEP TSA, 1990-1992 (m = 34) -- t, df = 60.
Compares unadjusted p-values with Bonferroni adjusted significance,
 Hochberg's (1988), and Benjamini & Hochberg's (1995) techniques.
*********************************************************************;

data tsa;
 t = (natdiff-meandiff)/pooled;
 p_value = 1-(probt(abs(t),60));
if p_value le (.05/2) then un = '*'; else un = ' ';
if p_value le (.05/68) then b1 = '*'; else b1 = ' ';
cards;
...;
proc sort data = tsa out = psort1; by p_value;

data compute; set psort1;
 i = _n_;
 p_bh = (i/34)*.025;
 p_b2 = (1/(34+1-i))*.025;
proc sort data = compute out = psort2; by descending p_value;

data do_bh; set psort2;
 retain value 0;
if value then goto seq;
if p_value le p_bh then value = 1;
   else do;
        value = 0;
        bh = ' ';
        return;
   end;
seq: bh = '*';
 drop value;
proc sort; by state;

data do_b2; set psort2;
 retain value 0;
if value then goto seq;
if p_value le p_b2 then value = 1;
   else do;
        value = 0;
        b2 = ' ';
        return;
   end;
seq: b2 = '*';
 drop value;
proc sort; by state;

data print; merge do_bh do_b2; by state;
proc sort; by p_value;
proc print; var state meandiff pooled t p_value
                un b1 p_bh bh p_b2 b2;
run;
```